(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 June 2003 (19.06.2003)

PCT

(10) International Publication Number
WO 03/050708 A1

(51) International Patent Classification[7]: G06F 15/173

(21) International Application Number: PCT/US02/38687

(22) International Filing Date: 4 December 2002 (04.12.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/341,098 7 December 2001 (07.12.2001) US

(71) Applicant *(for all designated States except US)*: **VITESSE SEMICONDUCTOR COMPANY** [US/US]; 741 Calle Plano, Camarillo, CA 93012 (US).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: **ELHOJ, Martin** [DK/SE]; Sundspromenaden 25, S-211 16 Malmo (SE).

(74) Agent: **CASTELLANO, John, A.**; Harness, Dickey & Pierce, P.L.C., P.O. Box 8910, Reston, VA 20195 (US).

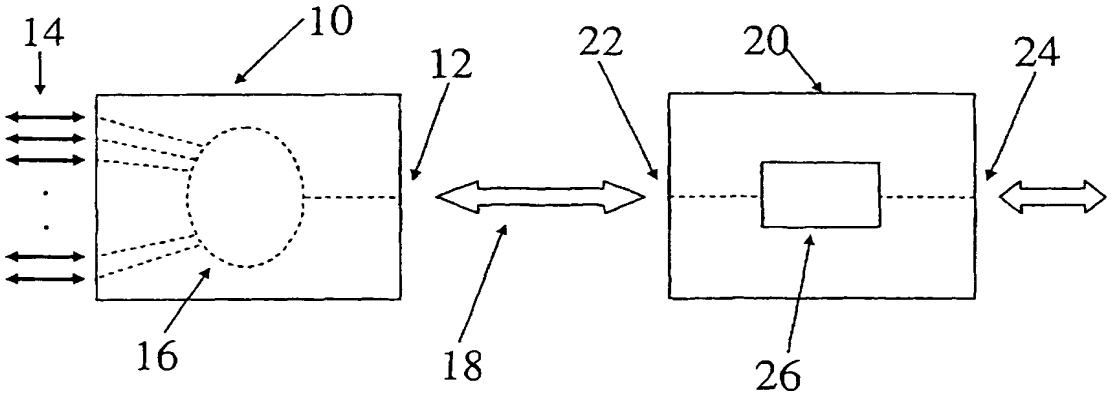
(81) Designated States *(national)*: AE, AG, AL, AM, AT (utility model), AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ (utility model), CZ, DE (utility model), DE, DK (utility model), DK, DM, DZ, EC, EE (utility model), EE, ES, FI (utility model), FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK (utility model), SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) **Title:** A METHOD AND SYSTEM FOR TRANSPORTING INFORMATION VIA THE XGMII OR FIBER CHANNEL STANDARD



(57) **Abstract:** Information is transformed between two link partners (10,20) as Sequence Ordered Sets or Signal Ordered Sets in XGMII or Fibre Channel (18). Such sets are defined to not be amended by intervening PHY's and may be transmitted even when there is no other traffic between the link partners. The information transmitted may be congestion information in a situation where there is heavy traffic in one direction between the link partners (10,20) and little or no traffic in the other direction.

WO 03/050708 A1

A METHOD AND SYSTEM FOR TRANSPORTING INFORMATION VIA THE XGMII or FIBER CHANNEL STANDARD

The present application hereby claims priority under 35 U.S.C. §119 on US patent publication number 60/341,098 filed December 7, 2001, the entire contents of which are hereby incorporated by reference.

The present invention relates to a method and a system for transmitting information via the XGMII standard as defined in IEEE 802.3ae clause 46 or Fibre Channel (10GFC) standard, which presently refers to the XGMII standard.

The XGMII standard presently is used for relaying data between network elements and defines that state or condition information may be transmitted between link partners only when one of these fails. The XGMII standard also defines that no matter how the XGMII signals are transmitted, such as via XAUI, the intervening elements, such as PCS's/PMA's/PMD's – or assembled into a PHY, must acknowledge and relay specific XGMII control information without amendment.

The present invention relates to the use of the XGMII ability to transmit control signals through transmission channels while ensuring that the signals actually reach the recipient. However, the invention relates to the use of that ability as a part of a normal operation of the link partners – not a failure mode.

The Fibre Channel standard has adopted the ability and functionality of the XGMII standard.

In EP-A-1 133 124 it may be seen how link partners (PHY's) communicate on top of a standard XGMII communication. However, such PHY's still have to transmit XGMII control signals un-amended. The information added on top of the XGMII information is removed before presenting the XGMII signals to the network element. These signals are used merely for the operation of the PHY's.

Thus, in a first embodiment, the invention relates to a method of transporting information from a first network unit to a second network unit, the method comprising:

- providing the network units being adapted to communicate via the XGMII or Fibre Channel standard defining a simultaneous transfer of a number of words of information relating to data or control information, each network unit being adapted to determine whether information received relates to data or control information, the network units communicating via a transmission path,
- providing, in the first network unit, condition information relating to one of a number of conditions obtainable in the first network unit during normal operation,
- transmitting, over the transmission path as control information, the condition information to the second network unit.

The XGMII standard may be seen from IEEE Draft P802.3ae/D3.2 where it is clear that the standard, at present, defines that four lanes of 8 bits (one word) transmit a total of four words simultaneously – at 156,25 MHz DDR giving a total of 10 Gbit/s of data transfer. In addition to that, each XGMII interface lane comprises a control signal conductor per lane - and the interface has a clock.

In the present context, the control information will be a full lane of information meaning that the control information will be transmitted (maybe together with other information) in the number of words defined to be transmitted simultaneously by the XGMII/FC standard.

The XGMII standard additionally defines that four simultaneously transmitted words (one in each of the four lanes) may relate to data to be transmitted or to control information. The control information may be transmitted in the form of so-called Sequence Ordered_Sets (SOS) which are defined to have the first word (in lane 0) represent "9C hex" and which are presently only defined to signal (by the contents in the other three words/lanes) that either the local or the remote link partner fails. Thus, the SOS' are only used in non-operational situations.

In Fibre Channel (FC), the corresponding feature is the so-called Signal Ordered_Sets, having a header of "5C hex". These SOS' are also possible in XGMII.

It should be noted that XGMII signals may be transmitted as defined over small distances or over communication channels based on other communication standards, such as

XAUI, 10GBASE-SR, 10GBASE-SW, 10GBASE-LX4, 10GBASE-LR, 10GBASE-LW, 10GBASE-ER, or 10GBASE-EW , which may be transmitted over larger distances. When using the other standards, normally PHY's are used for performing this conversion. A XAUI PHY may, for timing reasons, delete one of two consecutive SOS'es but must transmit single SOS'es. Transmitted SOS'es must be transmitted un-amended. PHY's may, naturally be positioned on the same chip as the networking element with which it communicates via an XGMII interface. In that manner, the XGMII is only an internal interface.

It should be noted that, in contrast with the XGMII and FC standards, at present, "data" is information relating to the data packets which are desired transported over the interface or transmission path. On the other hand, "control information" does not carry data. Thus, a SOS (which according to the standard has a Control header word and where the information carried on the other three lanes would be denoted data) will in the present context be denoted to consist only of control information.

In the present context, "normal operation" is a non-fault operation relating to an operation where the network element operates (i.e. transmits and receives data – if there is any to receive) as intended, such as where a condition is reported, such as:

- congestion,
- over heating,
- a change in mode of operation,
- information as to which of a plurality of modes of operation is presently used,
- information relating to the amount, quality, priority or the like of data received, processed, output or otherwise handled at a networking element.

Information as to this condition is reported to the other networking element, which may then take appropriate action. Such action could be to alter a present mode or manner of operation in accordance with the condition of the reporting network element. Such alteration could be to inform an operator of the condition of the first network element – or to e.g. reduce a data rate of data transmitted to the first element.

In a preferred embodiment, the second network unit transmits data to the first network unit, and the second network unit alters an actual mode of operation on the basis of the condition information received from the first network unit.

Also, in the providing step, the first network unit could provide information relating to congestion therein. Then, the second network unit preferably reduces or stops the transmission of data to the first network unit upon receipt of the condition information.

When the first network unit receives the data from the second network unit and outputs the data on a plurality of output ports preferably, in the providing step, the first network unit provides condition information relating to one or more of the output ports being congested. In that situation, the second network unit could reduce or stop only data transmission to the port(s) congested and not to the ports, which are not congested.

Alternatively or additionally, when the first network unit receives the data from the second network unit and outputs the data on a plurality of output ports then the first network unit could monitor or determine a bandwidth utilization of each output port, and provide condition information relating to which of the output ports should receive less or no data. Also in this situation, the second network unit may adapt transmission of data to the individual ports on the basis of the information received.

Thus, the second network unit may reduce or stop transmission of data to the identified port(s) of the first network unit upon receipt of the condition information.

The amount of information, which may be transferred in a single SOS, will be limited by the size of the three words (when one word is reserved for the identification of the SOS). If more information is needed, the condition information may be divided into a number of parts where each part is transmitted in a separate SOS. A bit map or other information may be provided in each SOS in order for the receiving element to be able to identify which part of the information is received and then to regenerate the information transmitted. In this manner, any amount of information may be transferred. However, due to the standard specifying that any e.g. PHY may delete one of two consecutive SOS'es – and up to one Idle, it is desired that such SOS'es are not transmitted consecutively and are furthermore spaced by sufficiently many Idles to not suddenly become "neighbours". Thus, preferably, the transmitting step comprises transmitting the condition information as a plurality of first Sequence Ordered_Sets or Signal Ordered_Sets, where one or more numbers of words relating to data or other control information are transmitted between each pair of the plurality of first Sequence Ordered_Sets or Signal Ordered_Sets.

Preferably, Idle's are used to at least be imminent to these SOS'es. Also, preferably at least as many Idles are used as there are PHY's in/at the communication path.

In the preferred embodiment, the first network unit transmits data packets to the second network unit via the communication path, and provides the condition information as control information only when no data information has been transmitted for a predetermined period of time.

In this embodiment, the first network unit preferably also transmits condition information to the second network unit as part of a data packet transmitted as data over the communication path. In this manner, the information is transmitted as part of data packets and only as control information when no data packets are transmitted. This ensures that the information is transmitted no matter whether data is transmitted or not.

In that situation, the condition information is preferably transmitted at predetermined time intervals as long as no data is transmitted. These time intervals may be determined in a number of ways. Presently, a time interval corresponding to the time it takes to transmit ½ maximum data packet is preferred. If information is received at the second network unit both from a data packet and as control information within a predetermined time period, the information from a predetermined one of the data packet and the control information is used and the information from the other is discarded. This is mainly due to the fact that the circuits deriving this information will normally differ – and so will the latencies thereof.

In the present context, a "network unit" may be any unit or part thereof communicating XGMII information with another unit or part thereof. Thus, a "unit" may be a complete switch, aggregator, router, or be part thereof, such as a packet or frame processor, a network processor, a storage medium or the like. A unit will normally be a single chip or a combination of chips.

It should be noted that the "pure" XGMII communication may be solely an internal interface of a chip or be used as a chip-to-chip communication. For longer distance communications other standards, such as XAUI, may be used for transporting the XGMII information. In the latter case, the communication path will also comprise e.g. a pair of PHYs – one for each end of an optical fibre for receiving the "pure" XGMII information and for converting it into a XAUI signal before transmission along the optical fibre (the XAUI

standard specifies a single optical fibre carrying four wavelengths – even though some proprietary solutions use multiple fibres). Naturally, these PHYs may form part of a chip incorporating part of or all of the pertaining network element.

The first and/or second network unit may additionally perform the operation of a switch, an analyser, a packet processor, a hub, a router, an aggregator, a deaggregator, a multiplexer, or a demultiplexer. These circuits may communicate using any standard supporting or acknowledging the XGMII standard.

Also, in one situation, the communication channel comprises a number of parallel conductors and wherein the network units communicate via the number of parallel conductors so as to, at least substantially simultaneously on the conductors, transfer the number of words. In this situation, the first and second network units preferably transmit, at least simultaneously and on at least part of the parallel conductors, a predetermined number of bits, such as 1 bit. This interface is normally used for small distances, such as inter-chip connections or connections within a single chip.

In another situation, the communication along the communication channel is additionally performed in accordance with the XAUI standard. This communication may take place both over electrical wires or via optical fibres. Thus, the XGMII information may, in accordance with the invention, be converted into other standards as long as these acknowledge the integrity of the XGMII control information.

Normally, the communication path comprises a pair of PHYs receiving the information to be communicated on a physical medium (optical fibre, wires or over the air) and/or receiving the information communicated from the medium, and which are adapted to transmit and receive the information to and from the medium without amending any XGMII control information thereby. As is seen in EP-A-1,133,124, the PHY's may add information into the XGMII data stream. This additional information is, though, removed by the receiving PHY before the actual XGMII data stream is regenerated.

In general, the words defined to be simultaneously transmitted may constitute an XGMII Sequence Ordered_Set or Signal Ordered_Set.

In a second aspect, the invention relates to a system comprising:

- a first network unit and a second network unit adapted to communicate via the XGMII or Fibre Channel standard defining a simultaneous transfer of a number of words of information relating to data or control information, each network unit being adapted to determine whether information received relates to data or control information, the network units communicating via a transmission path,
- the first network unit being adapted to provide condition information relating to one of a number of conditions obtainable in the first network unit during normal operation,
- the first network unit being adapted to transmit, over the transmission path as control information, the condition information to the second network unit.

In the present context, when a unit or the like is adapted to perform an action or step, this unit or the like will have means for performing the action or step.

Preferably, the second network unit is adapted to transmit data to the first network unit, and wherein the second network unit is adapted to alter an actual mode of operation on the basis of the condition information received from the first network unit. Then, the first network unit could be adapted to provide condition information relating to congestion therein. Also, the second network unit could be adapted to reduce or stop the transmission of data to the first network unit upon receipt of the condition information.

Additionally or alternatively, the first network unit could comprise a number of output ports and could be adapted to receive the data from the second network unit and output the data on the plurality of output ports and wherein the first network unit could be adapted to provide condition information relating to one or more of the output ports being congested. When the first network unit comprises a number of output ports and is adapted to receive the data from the second network unit and output the data on the plurality of output ports, the first network unit could comprise means for monitoring or determining a bandwidth utilization of each output port, and means for providing condition information relating to which of the output ports should receive less or no data. Then, the second network unit could be adapted to reduce or stop transmission of data to the identified port(s) of the first network unit upon receipt of the condition information.

Preferably, the first network unit is adapted to transmit data packets to the second network unit via the communication path, and to provide the condition information as control information only when no data packets have been transmitted for a predetermined

period of time. Then, the first network unit could be adapted to also transmit condition information to the second network unit as part of a data packet transmitted as data over the parallel conductors. Also, the first network unit could comprise means for transmitting condition information at predetermined time intervals as long as no data is transmitted. Preferably, this time interval is that which it takes half a data packet of a maximum size to be communicated on the communication path.

The first and/or second network unit may be adapted to additionally perform the operation of a switch, an analyser, a packet processor, a hub, a router, an aggregator, a deaggregator, a multiplexer, or a demultiplexer.

As described above, the first network unit may be adapted to divide the condition information into a plurality of parts and transmit each part in one of a plurality of first Sequence Ordered_Set or a Signal Ordered_Set and to transmit one or more words relating to data or other control information between each one of the plurality of first Sequence Ordered_Sets or Signal Ordered_Sets.

In one situation, the communication channel comprises a number of parallel conductors and wherein the network units are adapted to communicate via the number of parallel conductors so as to, at least substantially simultaneously on the conductors, transfer the number of words. Then, the first and second network units may each be adapted to transmit, at least simultaneously and on at least part of the parallel conductors, a predetermined number of bits, such as 1 bit.

In another situation, the first and second network units may be adapted to communicate along the communication channel in accordance with the XAUI standard. Naturally, the communication between the first and second network units may take place along a number of different standards, such as firstly on a standard XGMII interface to another element, which then converts the signals into XAUI signalling and forwards these signals to the other network element – optionally via additional elements. This may be achieved by the communication path comprising a pair of PHY's receiving the XGMII information for transmission on a physical medium and/or receiving the information from the medium, and which is adapted to transmit and receive the information without amending any of the present XGMII control information thereby.

In general, the words defined to be transmitted simultaneously may constitute an XGMII/Fibre Channel Sequence Ordered_Set or a Signal Ordered_Set.

In a third aspect, the invention relates to an XGMII or Fibre Channel bit stream transmitted from a first network unit to a second network unit, the bit stream comprising four parallel transmissions each of one word of information, each word of information relating to either data or control information, wherein

- in each of the four parallel transmissions, no data is transmitted for a first predetermined period of time, and
- during the first predetermined period of time, one or more Sequence Ordered_Sets or Signal Ordered_Sets comprising condition information relating to one of a number of conditions obtainable in the first network unit during normal operation are transmitted with an interval of a second predetermined period of time.

Again, the condition information may relate to a congestion condition of the first networking unit.

Naturally, this bit stream may be e.g. a XAUI bit stream incorporating therein the XGMII information of the present type.

A fourth aspect of the invention relates to a method of transporting information from a first network unit to a second network unit, the method comprising:

- providing the network units being adapted to communicate via the XGMII or Fibre Channel standard defining a simultaneous transfer of a number of words of information relating to data or control information, each network unit being adapted to determine whether information received relates to data or control information, the network units communicating via a transmission path,
- providing, in the first network unit, information to be transported, and
- transmitting, over the transmission path as control information, the information to the second network unit.

This additional information may not relate to the networking element itself but may form part of information transmitted over the networking elements in a side band channel. This information may relate to the data packets transmitted or may be for use in subsequent networking units for management thereof. Such information is transmitted as control information in order to not strain the bandwidth of the communication link with data packets or frames which tend to take up more bandwidth than actually required.

The method may further comprise receiving the information, dividing the information into a plurality of information parts, and transmitting the information parts as individual control information to the second networking unit. In this manner, even larger portions of data – or continuous streams of data may be handled by simply dividing them and transmitting the parts individually. In this situation, the method preferably also comprises the second networking unit combining the individual information parts before outputting.

In fact, the method may comprise receiving, at the first networking unit, a data packet or frame, providing information relating to the packet or frame, transmitting the packet or frame to the second networking unit as data information, and transmitting the provided information to the second networking unit as control information. In this manner, the second networking unit receives information relating to the packet – but transmitted not in the packet.

Naturally, this additional information may be provided in packet preambles and SOS'es (control information) depending on the situation. SOS'es may be used when not enough preambles or packets are available – or simply to save bandwidth on the communication channel.

In the following, preferred embodiments will be described with relation to the drawing wherein:

- Fig. 1 illustrates a box diagram of two network units communicating with each other,
- Fig. 2 illustrates an Ethernet packet preamble and its transmission on an XGMII interface,
- Fig. 3 illustrates in more detail the box diagram of Fig. 1,
- Fig. 4 illustrates how condition information is transmitted in an XGMII SOS,

- Fig. 5 illustrates an XGMII bit stream, and
- Fig. 6 illustrates an embodiment alternative to that of Fig. 3.

The following description is limited to the use of Sequence Ordered_Sets in XGMII. It should be noted that the implementation of the invention in XGMII or Fibre Channel using Signal Ordered_Sets would be quite similar to the present description.

In Fig. 1, two network units, 10 and 20 are illustrated communicating via a communication channel 18. The unit 10 comprises one high throughput I/O 12 and 10 lower throughput I/O's 14. Also, the unit 10 comprises means 16 for aggregating/de-aggregating between the inputs 12 and 14.

The unit 20 has two high throughput I/O's 22 and 24 and a frame analysing engine 26 adapted to receive a data packet from the I/O 24, analyse it and alter the preamble of the packet with the findings of the analysis before transmitting the packet to the I/O 22 and the unit 10.

The result of the analysis of analyser 26 is incorporated into the preamble of the packet. The packet – including the preamble – is transmitted to the element 10, where the preamble information is used in the means 16 for determining which of the I/O's 14 should receive and output the packet.

The advantage of adding the findings of the analysis in the preamble of the packet is that this is transmitted with the packet at any rate and that this information is then transmitted without requiring any additional bandwidth.

This preamble may be used in a number of additional situations, one being the situation where one of the I/O's 14 is congested. In this manner, it might be desired to actually have the unit 20 reduce or stop the flow of packets to that particular I/O.

Thus, this congestion information may be transmitted in packets received by unit 10 and transmitted to the unit 20 – even though this information has no relevance to the packets wherein it is transmitted.

In that manner, a backpressure mechanism is obtained without requiring any additional bandwidth on the channel 18.

However, this backpressure mechanism only functions if packets are actually transmitted from the unit 10 to the unit 20. In the situation where no packets are transmitted in that direction but the data traffic in the opposite direction is large enough for one of the I/O's 14 to congest, that backpressure mechanism does not work.

In order to handle this situation, the so-called Sequence Ordered_Sets (SOS) of the XGMII standard are used. These SOS have a header of "9C hex". Alternatively, the Signal Ordered_Sets, having a header of "5C hex", could be used in exactly the same manner.

According to the invention, the communication on the channel 18 takes place via XGMII, which defines SOS as being a specific manner of transporting error information from one link partner to the other. A SOS is a column (4 words) where the first word, or lane 0, defines a specific control character). Such a SOS must be transmitted un-amended by any intervening networking equipment or parts such as PCS'es (PHY's etc) independently on which physical layer (XAUI or others) is used in the PHY's.

Presently, a SOS is only defined for use when one of the two link partners actually fails. However SOS'es may, according to the invention, be used in a number of other situations, such as for informing one link partner of another link partner's congestion.

Thus, in the situation where no traffic exists in the direction where the congestion information needs be transferred, a SOS is used having the defined word in Lane 0 but which now comprises information as to congestion - such as which of the I/O's (if more than one is present) is congested.

If no data are transferred, the use of a SOS avoids the requirement for transmitting idle or empty packages to transmit the congestion information over the channel 18. Standard Ethernet Pause Frames may be used, but these 64 byte frames may take up bandwidth on the interface.

It should be noted that it is preferred to transmit the port state information no matter whether any changes have occurred. This means that all packets preferably have the port

state information and, if no packets are transmitted, SOS'es are transmitted with the desired interval. Thus, if a change in port status has happened but the first SOS is dropped by a PHY (this can happen) the next SOS (or packet preamble) will update the network element accordingly.

In that manner, the congestion information is transmitted over the channel 18 even when no data packets and preambles are transmitted.

In order to ensure that the receiving unit 10 or 20 receives sufficiently frequent updates on congestion when no data packets are transmitted, the transmitting unit will transmit a SOS with time intervals corresponding to ½ max frame. Thus, both units will be updated with reference to any congestion whether data is transmitted or not – and none of the methods take away bandwidth from the data transport on the channel 18.

Fig. 2 illustrates the structure of a preferred embodiment of an Ethernet packet preamble and the division thereof into the words for transmission on a XGMII interface.

The XGMII interface comprises, for data/control transport, four lanes of 8 conductors each running at 156,25 MHz, double data rate. Thus, 10 Gbit/s may be transmitted over that interface. Each lane also comprises an additional conductor, which informs the receiver of whether the data of the lane is data or control information. Control information may be Start Of Packet, End Of Packet, Error, or Idle. The four lanes carry, simultaneously, four words – or one column of information. When the XGMII information is transmitted over a physical medium, these words may be delayed in relation to each other, but standard PHY's will correct this before presenting the XGMII signals to the recipient.

The SOS is a specific type of column where the first lane (Lane 0) has a predetermined content defining a SOS. The remainder of the column is, in practise, reserved but only two values (representing remote fault or local fault) are used/defined.

In Fig. 2, it is seen that the preferred amended preamble (bottom) of an Ethernet data packet actually takes up the 7 "most significant" bytes of the standard preamble. The 8 bytes of the preamble are transferred over the XGMII interface (top) in two cycles. This preamble may comprise information relating to a number of things, such as outgoing

congestion information relating to one or more of the ports 14, information as to at which of the ports 14 the packet was received, special purpose routing information or the like.

Part of this preamble relates to congestion information from one unit 10 or 20 to the other.

An Ethernet data packet comprises the preamble, a header having destination/source/other information, and a payload portion. After transmitting the preamble, as described above, the remainder of the data packet is, naturally, transmitted over the interface.

Fig. 3 illustrates a more detailed block diagram describing further elements in the units 10 and 20. Even though the signalling method may be used in a number of different systems, it is described between an aggregator/deaggregator 10 and an analyser 20.

The unit 10 has an I/O 12 now are illustrated to have an aggregator/deaggregator 16 and the I/O's 14 now represented by MAC/'sPHY's (which are illustrated as on-chip elements but which may as well be off-chip elements).

The communication path has been illustrated to have two PHY's 12' and 22' which are illustrated as off-chip PHY's but which may as well be on-chip PHY's. The elements 10 and 20 communicate with the respective PHY's using a standard XGMII interface.

The presently preferred PHY's are standard XAUI PHY's communicating via an optical fibre.

The unit 20 still has its I/O 24 now represented by an on-chip PHY and the analyser 26. This unit now also has a queuing means 28 for holding data packets having been analysed by the analyser 26 and before transmission to the unit 10. This queuing means 28 has one queue for each of the I/O's 14 of the unit 10. Naturally, the means 28 may optionally have more than one queue for each I/O 14 - such as one per priority.

Data packets are transmitted from the queuing means 28 to the unit 10 on a round robin basis, unless congestion information from the unit 10 informs the unit 20 of holding back packets for a defined one (or more) of the I/O's 14. In that situation, this/these queue(s) in the queuing means 28 is/are skipped during the round robin.

Data packets received in unit 20 from the unit 10 are not analysed but merely relayed through the unit.

Also, it is seen that the high throughput I/O's 12, 22, and 24 are expanded to also illustrate Media Access Controllers 17 and XGMII Reconciliation Sub-layers 15. The MAC's 17 define the overall protocol for the communication on the link 18, which is now illustrated as two PHY's communicating via an optical fibre, and the MAC's strip away the preamble information transmitted as part of a packet. The RS's 15 are the ones introducing and stripping off the SOS'es and which will then provide any additional information provided therein – such as the present congestion information.

Naturally, congestion information may also be provided from the unit 20 to the unit 10. In that situation, congestion information in any of the port state SOS or preamble bits used for signalling congestion would stop or reduce the data flow from the unit 10 to the unit 20.

In Fig. 3 is also illustrated two additional elements 50 defining an additional route of information which may be transmitted between the network units. The information transmitted may be any information – and even information not relating to neither the actual networking unit or the packets handled thereby. This information may be network managing information or fully independent information desired transported over the network.

This information may be provided as a stream of data or as individual packets or frames which may be transmitted as normal packets or frames if there is bandwidth on the channel 18. The information may be transported in preambles of the packets transported or may be transported in SOS'es if there are not enough packets or preambles available – or simply if desired. In this situation, the below minimum time or clock distance between the SOS'es needs not be fulfilled in that this data may not be timing critical.

Fig. 4 illustrates how the port state information of the incoming Ethernet data packet only takes up 2 bytes of the port state SOS. In that manner, the information may be transmitted in a single column on the XGMII interface. If more information is to be sent, it may be split up into a number of parts each transmitted in a port state SOS. Reassembly of the information may be performed in any applicable manner. It should be noted that as a PHY

may delete one of two consecutive SOS'es and one Idle column, the port state SOS'es should not be transmitted next to each other. They should be separated by a number of Idles corresponding to at least the number of PHY's between the elements.

It is important not to loose any of the present port state SOS columns. Thus, it is ensured that such SOS columns are transmitted one at the time, due to the following definition IEEE P802.3ae/D3.2 draft standard stating that (ref. Clause 48.2.4.2.3):

- "Sequence Ordered_Sets may be deleted to adapt between clock rates;
- Sequence Ordered_Set deletion occurs only when two consecutive sequence Ordered_Sets have been received and deletes only one of the two;"

Thus, any given PCS will guarantee that a single (port state) SOS column is not deleted during error-free operation. However, due to the fact that the port state is transmitted irrespective of any changes, the element will only be "out of sync" shortly upon loss of a port state SOS.

It is seen that the SOS also incorporates an ID. Presently, the ID is 1 or 2, but as the standard is not yet fixed, it is preferred that both the ID, the SEQ and the condition information – that is, both the values thereof as well as their positions within the four lanes – be software defined so that any changes between the present state of the draft and the final standard may be taken into account without having to manufacture a new chip.

To comply with IEEE P802.3ae/D3.2 Sequence Ordered Set format, the word in Lane 3 should be ≥3. This value is preferably software programmable in order to be able to adapt to any amendments to this standard.

The information in the next two words/lanes may be used for representing congestion or not in up to 16 individual channels or I/O's of the unit 10. The final word is the SEQ, which has a defined value of 9C hex.

A guard band, consisting of 3 idle columns is added to each frame. A fully loaded 10 GB/s Ethernet line will always have at least two full IDLE columns as Inter Packet Gap. That number of IDLE columns is sufficient to ensure proper operation of the PHY components.

Setting the guard band to three columns will provide more IDLE columns thus still guaranteeing proper PHY operation even when a SOS column is inserted.

This is seen from Fig. 5, which illustrates different situations, which may occur in an XGMII bit stream. The bit stream is illustrated for each of the four lanes.

Each packet is transmitted with the Start of Packet (S) in the first lane, followed by two Port Status (PS) words and the remainder of the packet Preamble (P). Then, the Data (D) of the packet is transmitted followed by an End of Packet (E). Naturally, the full packet may end in any lane, and the remainder of the lane is filled with Idles (I).

It is seen that a guard band is inserted between two consecutive packets. Also, if no new packet is to be sent after transmission of a packet, the guard band is transmitted and then a SOS column. If still no packets are to be transmitted, a period of time (corresponding to "Space") corresponding to ½ maximum packet size is waited and another SOS column is transmitted. If a SOS column is scheduled within a packet, the SOS column will be transmitted (unless a new packet is to be transmitted) after the guard band following the actual packet.

Also, preferably, the contents of a port state SOS is always the latest information. Even though a port state SOS may be scheduled a period of time in advance of its transmission, the port state information is not added before immediately before transmission.

If the PHY's 12 and 22 are XAUI PCS'es, the requirement to have 3 IDLE columns will guarantee that the PCS has sufficient idle columns that can be converted to the /A/, /K/, and /R/ symbols required for proper operation.

It should be noted that as the port status information is derived (see Fig. 3) by the RS 15 from a SOS and the preamble information by the MAC 17, a problem might be encountered when the RS derives congestion information from a column following a data packet presently analysed by the MAC. Then the MAC may actually output congestion information later than that from the port state SOS even thought it was transmitted from the transmitting unit before the information in the port state SOS. Thus, presently, it is preferred that if both the RS and the MAC output information within a predetermined time

interval, only the information from the RS is used. This predetermined time interval will depend on the size of the guard band and the internal latency of the MAC.

Due to the fact that the present use of the XGMII standard may be intervened by the IEEE actually defining other use of the presently used SOS'es, the present manner of using the SOS'es is software programmable in the elements 10 and 20 in order to adapt already active elements when the IEEE alters the XGMII standard. Thus, Signal Ordered_Sets may be used instead or the actual manner of providing the status information in the SOS may be changed in order for the elements 10 and 20 to be able to be fully XGMII compliant and still have the present functionality.

Fig. 6 illustrates a system closely related to that of Fig. 3 where, however, compared to the element 20, the element 20' is rotated and has an additional functionality. In this embodiment, data from the element 10 is analysed in element 20' and transmitted to a switch 21. The preambles of the data packets are amended in element 20' so as to aid the switch 21 in determining from which port to output the data.

Data is also received by element 20' from the switch 21. This data is now not merely relayed through the element 20' (as was the case in the element 20) but is processed in a processing element 27. This processing is performed on the basis of information in the data preamble. This information is provided by the analysing element 26 which originally analysed the data before transmission to the switch 21.

This processing may be a processing which enlarges the data packet. By performing this type of processing, the bandwidth required at the switch is the same as that at required at the input of the element 20'. In that manner, no blocking will take place at the input of the analyser/switch system.

Naturally, the SOS port state information from the element 10 may be transmitted through the element 20' to the switch 21 in order for it to take this information into account when determining the switching order of data therein. It should be remembered that the preambles of data packets received from element 10 might comprise therein information as to which port the data packet was received at. The preambles output by the element 20' may additionally comprise information as to not only from which port of the switch 21 the packet should be output but also information for the outputting element 10 as to from

which port to output the data. This information may be used in the switch 21 in order to delay or drop packets for a congested port in the destination element 10. Alternatively, the packet should be delayed or dropped in the receiving element 20' or 10 – and thereby take up bandwidth along those elements.

5

CLAIMS

1. A method of transporting information from a first network unit to a second network unit, the method comprising:

- providing the network units being adapted to communicate via the XGMII or Fibre Channel standard defining a simultaneous transfer of a number of words of information relating to data or control information, each network unit being adapted to determine whether information received relates to data or control information, the network units communicating via a transmission path,
- providing, in the first network unit, condition information relating to one of a number of conditions obtainable in the first network unit during normal operation,
- transmitting, over the transmission path as control information, the condition information to the second network unit.

2. A method according to claim 1, wherein the second network unit transmits data to the first network unit, and wherein the second network unit alters an actual mode of operation on the basis of the condition information received from the first network unit.

3. A method according to claim 2, wherein, in the providing step, the first network unit provides information relating to congestion therein.

4. A method according to claim 3, wherein the second network unit reduces or stops the transmission of data to the first network unit upon receipt of the condition information.

5. A method according to claim 3, wherein the first network unit receives the data from the second network unit and outputs the data on a plurality of output ports and wherein, in the providing step, the first network unit provides condition information relating to one or more of the output ports being congested.

6. A method according to claim 2, wherein the first network unit receives the data from the second network unit and outputs the data on a plurality of output ports and wherein the first network unit monitors or determines a bandwidth utilization of each output port, and

provides condition information relating to which of the output ports should receive less or no data.

7. A method according to claim 5, wherein the second network unit reduces or stops transmission of data to the identified port(s) of the first network unit upon receipt of the condition information.

8. A method according to claim 1, wherein the first network unit transmits data packets to the second network unit via the communication path, and provides the condition information as control information only when no data information has been transmitted for a predetermined period of time.

9. A method according to claim 8, wherein the first network unit also transmits condition information to the second network unit as part of a data packet transmitted as data over the communication path.

10. A method according to claim 8, wherein the condition information is transmitted at predetermined time intervals as long as no data is transmitted.

11. A method according to claim 1, wherein the first and/or second network unit additionally performs the operation of a switch, an analyser, a packet processor, a hub, a router, an aggregator, a deaggregator, a multiplexer, or a demultiplexer.

12. A method according to claim 1, wherein each network unit determines, from one word of the number of words defined to be simultaneously transmitted, whether the information in the other words defined to be transmitted simultaneously therewith relate to data or control information.

13. A method according to claim 1, wherein the communication channel comprises a number of parallel conductors and wherein the network units communicate via the number of parallel conductors so as to, at least substantially simultaneously on the conductors, transfer the number of words.

14. A method according to claim 13, wherein the first and second network units transmit, at least simultaneously and on at least part of the parallel conductors, a predetermined number of bits, such as 1 bit.

15. A method according to claim 1, wherein the words defined to be simultaneously transmitted constitute an XGMII Sequence Ordered_Set or Signal Ordered_Set.

16. A method according to claim 1, wherein communication along the communication channel is additionally performed in accordance with the XAUI standard.

17. A method according to claim 1, wherein the communication channel comprises a pair of PHY's receiving the information for communication on a physical medium and/or receiving the information from the medium, and which are adapted to transmit and receive the information to and from the medium without amending any XGMII/Fibre Channel control information thereby.

18. A system comprising:

- a first network unit and a second network unit adapted to communicate via the XGMII or Fibre Channel standard defining a simultaneous transfer of a number of words of information relating to data or control information, each network unit being adapted to determine whether information received relates to data or control information, the network units communicating via a transmission path,
- the first network unit being adapted to provide condition information relating to one of a number of conditions obtainable in the first network unit during normal operation,
- the first network unit being adapted to transmit, over the transmission path as control information, the condition information to the second network unit.

19. A system according to claim 18, wherein the second network unit is adapted to transmit data to the first network unit, and wherein the second network unit is adapted to alter an actual mode of operation on the basis of the condition information received from the first networking unit.

20. A system according to claim 19, wherein the first network unit is adapted to provide condition information relating to congestion therein.

21. A system according to claim 20, wherein the second network unit is adapted to reduce or stop the transmission of data to the first network unit upon receipt of the condition information.

22. A system according to claim 20, wherein the first network unit comprises a number of output ports and is adapted to receive the data from the second network unit and output the data on the plurality of output ports and wherein the first network unit is adapted to provide condition information relating to one or more of the output ports being congested.

23. A system according to claim 19, wherein the first network unit comprises a number of output ports and is adapted to receive the data from the second network unit and output the data on the plurality of output ports and wherein the first network unit comprises means for monitoring or determining a bandwidth utilization of each output port, and means for providing condition information relating to which of the output ports should receive less or no data.

24. A system according to claim 22, wherein the second network unit is adapted to reduce or stop transmission of data to the identified port(s) of the first network unit upon receipt of the condition information.

25. A system according to claim 18, wherein the first network unit is adapted to transmit data packets to the second network unit via the communication path, and to provide the condition information as control information only when no data packets have been transmitted for a predetermined period of time.

26. A system according to claim 25, wherein the first network unit is adapted to also transmit condition information to the second network unit as part of a data packet transmitted as data over the parallel conductors.

27. A system according to claim 25, wherein the first network unit comprises means for transmitting condition information at predetermined time intervals as long as no data is transmitted.

28. A system according to claim 18, wherein the first and/or second network unit is/are adapted to additionally perform the operation of a switch, an analyser, a packet processor, a hub, a router, an aggregator, a deaggregator, a multiplexer, or a demultiplexer.

29. A system according to claim 18, wherein each network unit is adapted to determine, from one word of the number of words defined to be simultaneously transmitted, whether the information in the other words transmitted at least substantially simultaneously therewith relate to data or control information.

30. A system according to claim 18, wherein the communication channel comprises a number of parallel conductors and wherein the network units are adapted to communicate via the number of parallel conductors so as to, at least substantially simultaneously on the conductors, transfer the number of words.

31. A system according to claim 30, wherein the first and second network units are each adapted to transmit, at least simultaneously and on at least part of the parallel conductors, a predetermined number of bits, such as 1 bit.

32. A system according to claim 30, wherein the first and second network units are each adapted to transmit, in series, two parallel numbers of words representing control information.

33. A system according to claim 18, wherein the words defined to be simultaneously transmitted constitute an XGMII Sequence Ordered_Set or Signal Ordered_Set.

34. A system according to claim 18, wherein the communication channel is adapted receive, from each of the first and second network elements, the XGMII information, and to communicate along a physical medium in accordance with the XAUI standard.

35. A system according to claim 18, wherein the communication channel comprises a pair of PHY's receiving the information for to be communicated on a physical medium and/or receiving the information from the medium, and which are adapted to transmit and receive the information to and from the medium without amending any XGMII control information thereby.

36. An XGMII or Fibre Channel bit stream transmitted from a first network unit to a second network unit, the bit stream comprising four parallel transmissions each of one word of information, each word of information relating to either data or control information, wherein

- in each of the four parallel transmissions, no data is transmitted for a first predetermined period of time, and
- during the first predetermined period of time, two or more Sequence Ordered_Sets or Signal Ordered_Sets comprising condition information relating to one of a number of conditions obtainable in the first network unit during normal operation are transmitted with an interval of a second predetermined period of time.

37. A bit stream according to claim 36, wherein the condition information relates to a congestion condition of the first networking unit.

38. A method according to claim 1, wherein the transmitting step comprises transmitting the condition information as a plurality of first Sequence Ordered_Sets or Signal Ordered_Sets, where one or more numbers of words relating to data or other control information is/are transmitted between each one of the plurality of first Sequence Ordered_Sets or Signal Ordered_Sets.

39. A system according to claim 18, wherein the first network unit is adapted to divide the condition information into a plurality of parts and transmit each part in one of a plurality of first Sequence Ordered_Set or a Signal Ordered_Set and to transmit one or more words relating to data or other control information between each one of the plurality of first Sequence Ordered_Sets or Signal Ordered_Sets.

40. A method of transporting information from a first network unit to a second network unit, the method comprising:

- providing the network units being adapted to communicate via the XGMII or Fibre Channel standard defining a simultaneous transfer of a number of words of information relating to data or control information, each network unit being

adapted to determine whether information received relates to data or control information, the network units communicating via a transmission path,

- providing, in the first network unit, information to be transported, and
- transmitting, over the transmission path as control information, the information to the second network unit.

41. A method according to claim 40, further comprising receiving the information, dividing the information into a plurality of information parts, and transmitting the information parts as individual control information to the second networking unit.

42. A method according to claim 41, wherein the second networking unit combines the individual information parts before outputting.

43. A method according to claim 40, further comprising receiving, at the first networking unit, a data packet or frame, providing information relating to the packet or frame, transmitting the packet or frame to the second networking unit as data information, and transmitting the provided information to the second networking unit as control information.

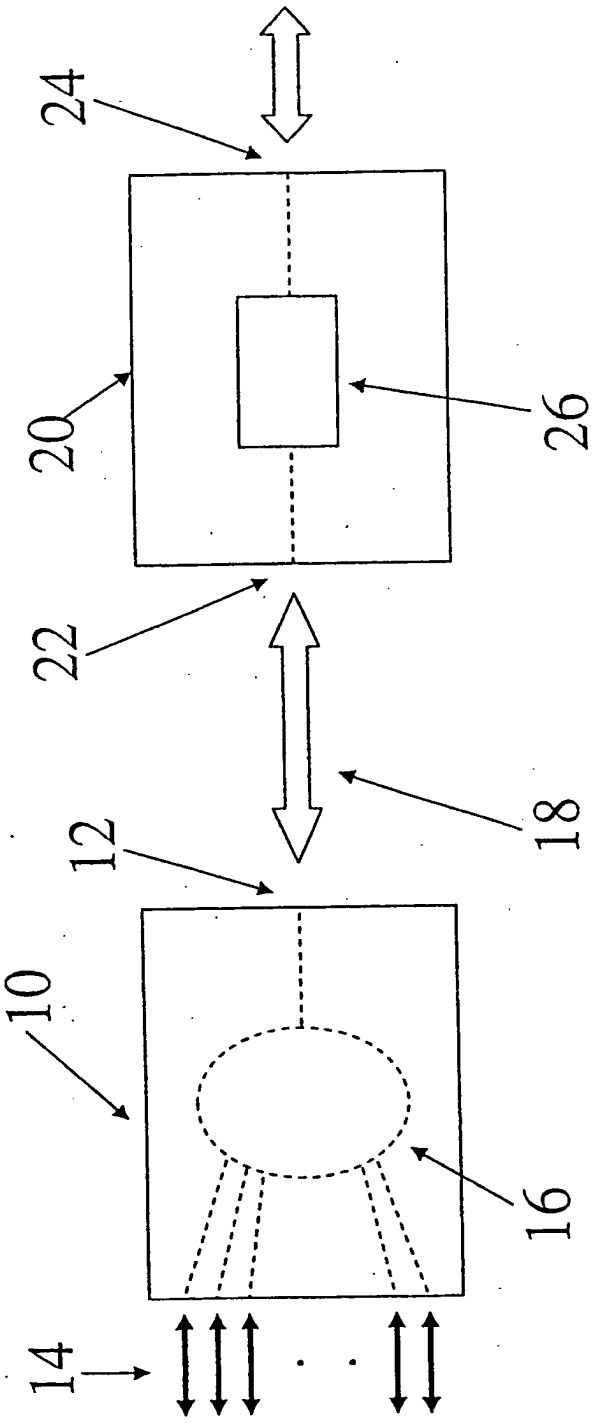


Figure 1

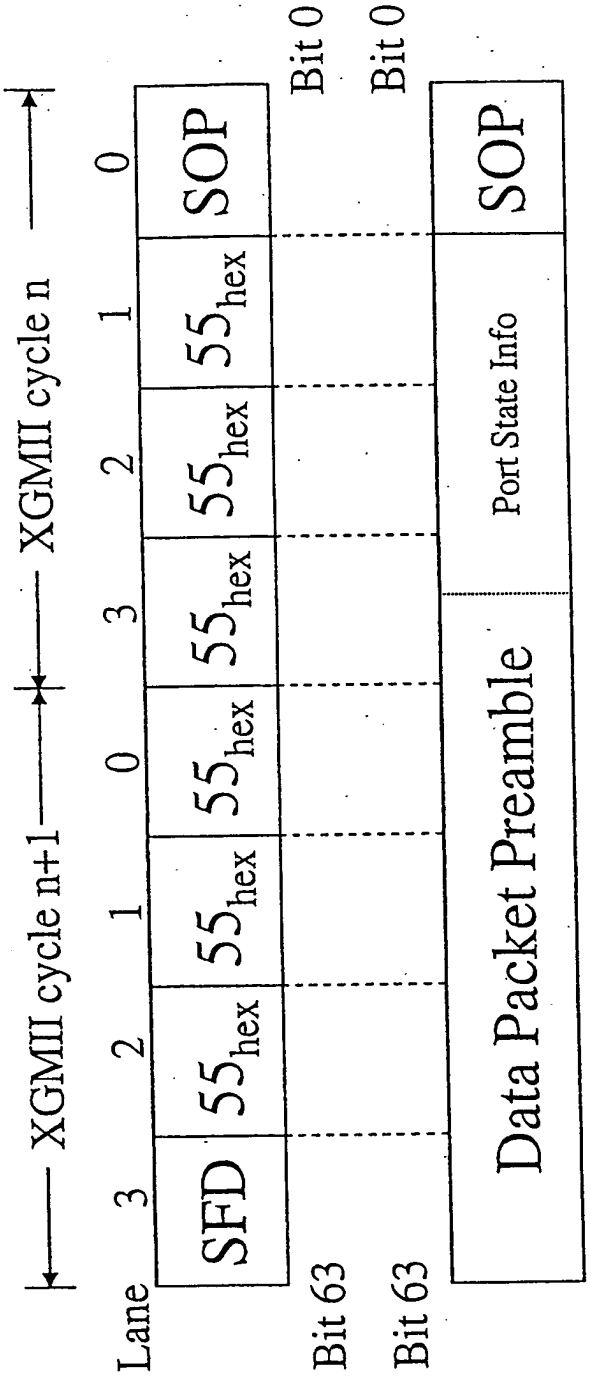


Figure 2

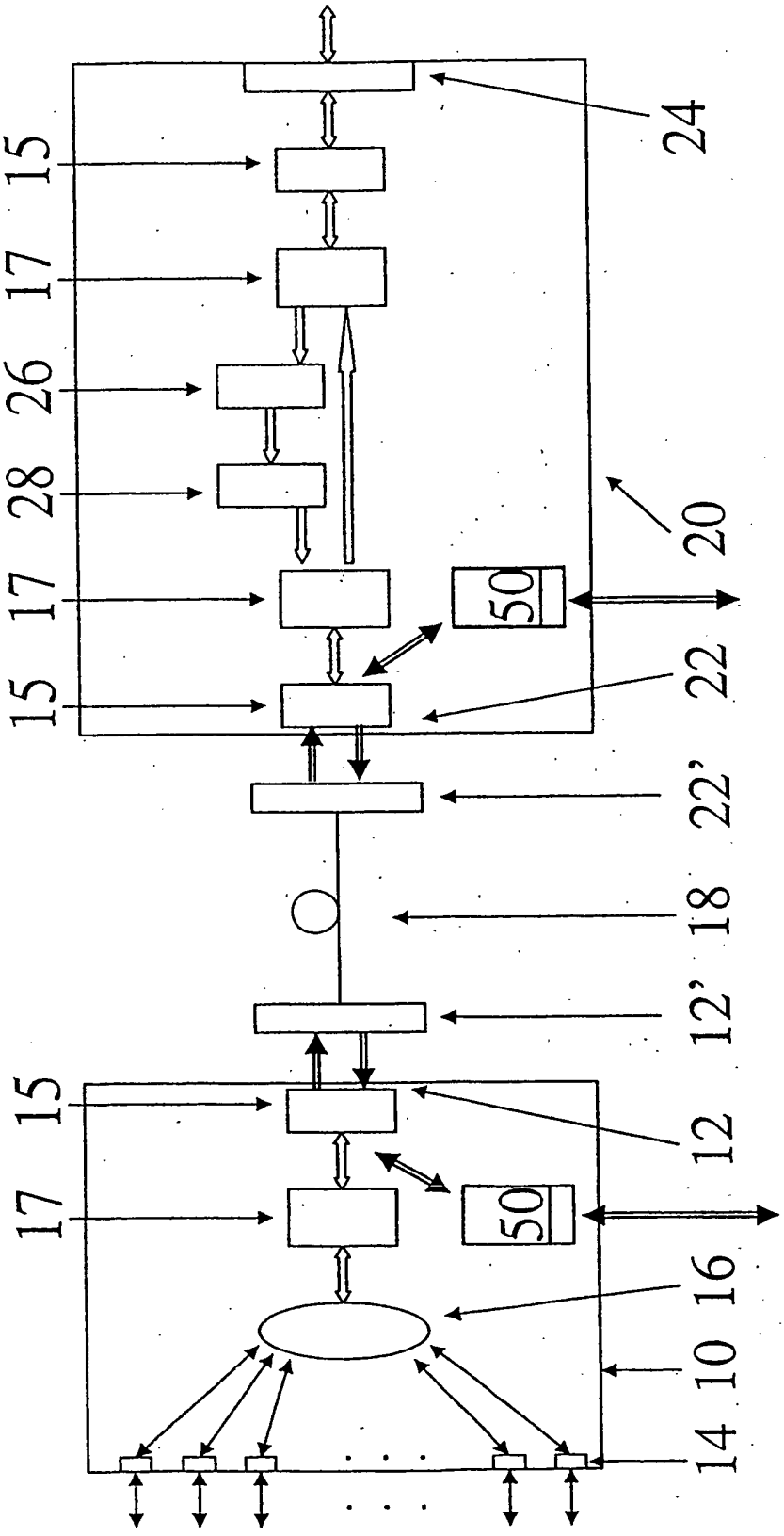


Figure 3

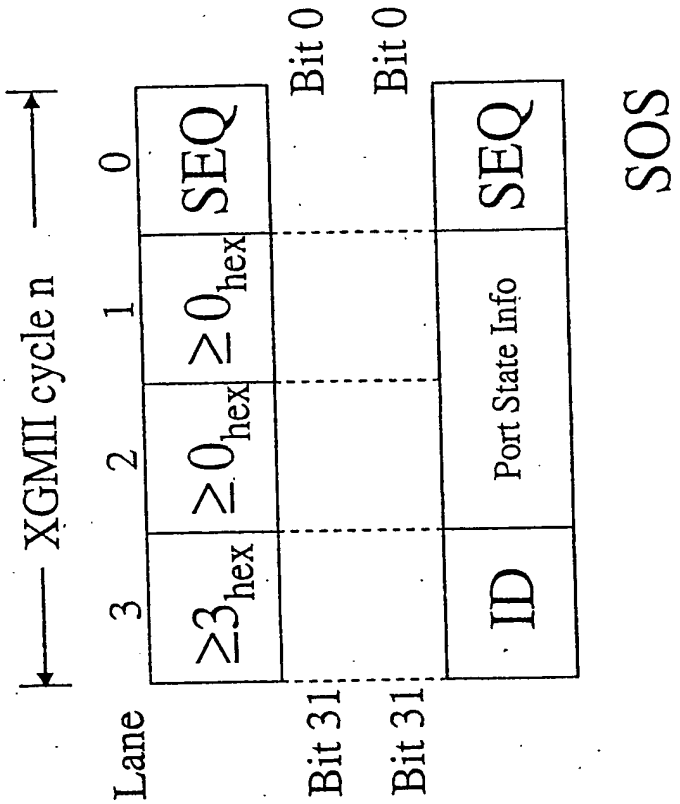

XGMII cycle n
Lane
3
2
1
0
Bit 31
Bit 31
Bit 0
Bit 0
≥3hex
≥0hex
≥0hex
SEQ
ID
Port State Info
SEQ
SOS

Figure 4

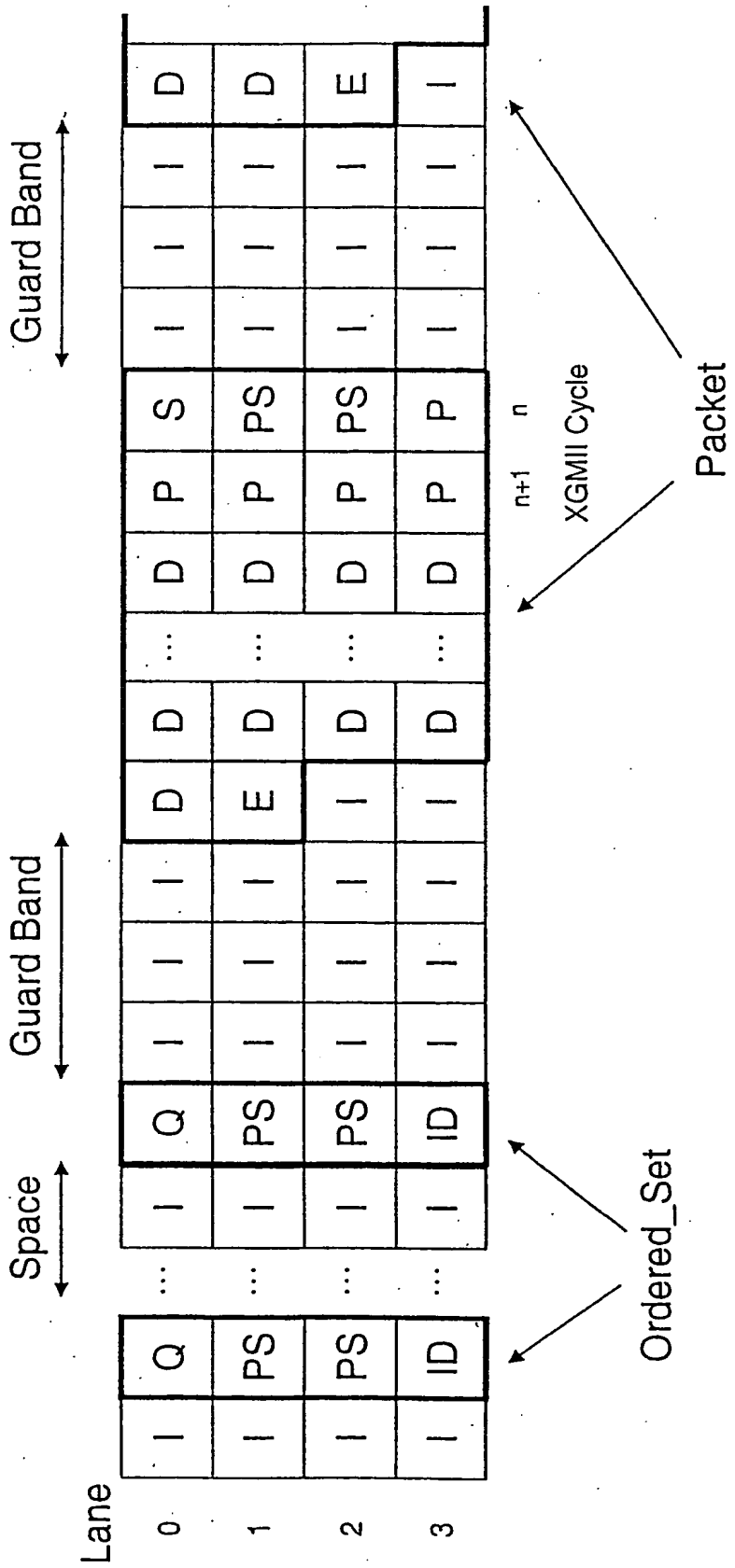

Lane
0
1
2
3
Space
Guard Band
Guard Band
Guard Band
Ordered_Set
Packet
XGMII Cycle
n+1
n
Q
PS
ID
S
P
D
E
I

Figure 5

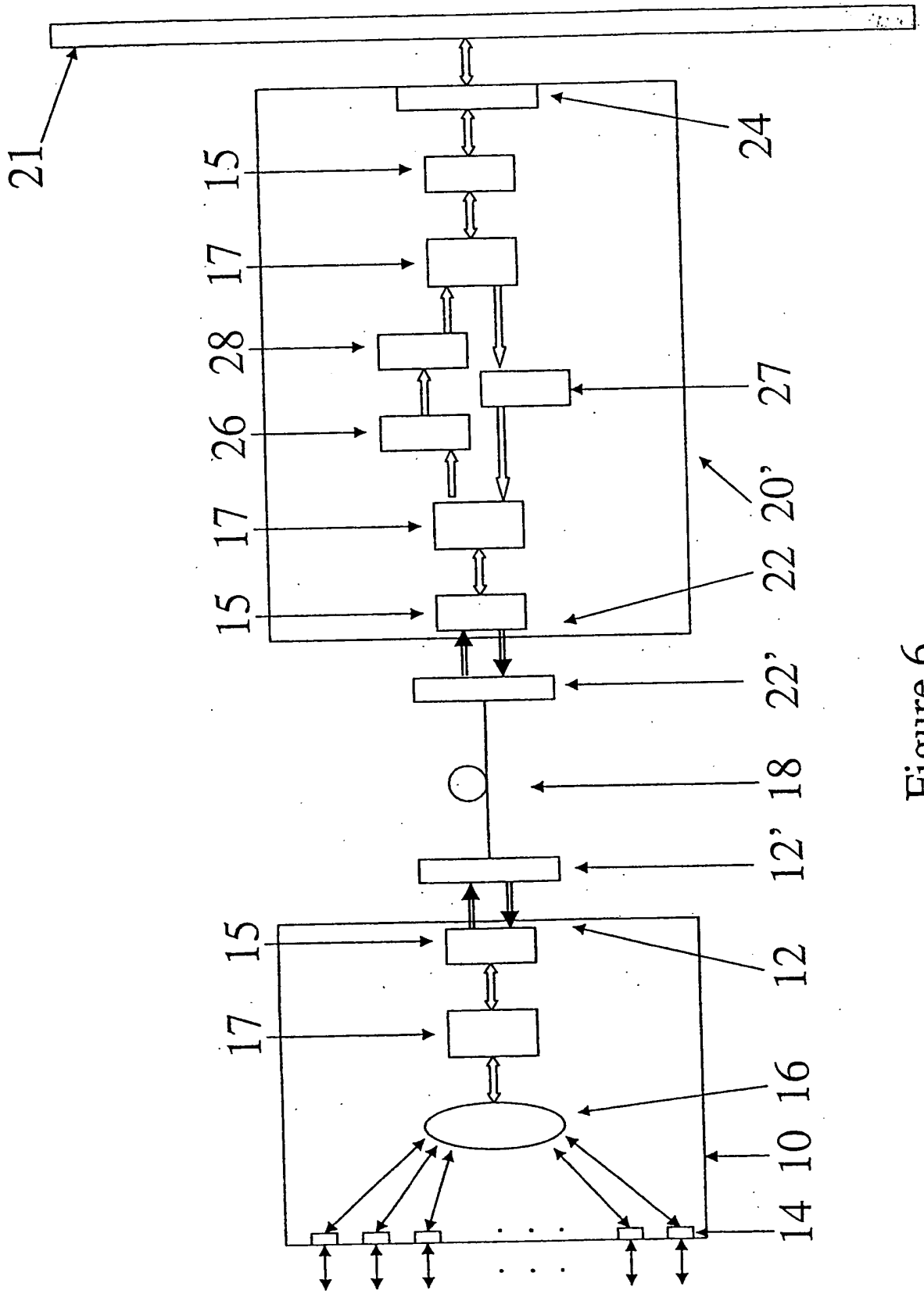


Figure 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US02/38687

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) :G06F 15/173
US CL :709/224

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 709/223, 232; 710/29, 33, 36, 40;

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

USPTO BRS: EAST
search terms: SGMII, Fibre chanel, netwrok, transfer, control data, links, PHY, condition, requirement.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y, P	US 2001/0034729 A1 (AZADET et al.) 25 October 2001, page 1, para. [0007]-[0009]	1-43
Y, P	US 2002/0172205 (TAGORE-BRAGE et al.) 21 November 2002, pages 2-4, pars. [0022]-[0089]	1-43
Y, E	US 2002/0194415 A1 (LINDSAY et al.) 19 December 2002, pages 2-3, pages [0017]-[0020]	1-43
Y, E	US 2003/0033463 A1 (GARNETT et al.) 13 February 2003, pages 1, pars.[0004]-[0013].	1-43

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 09 MARCH 2003	Date of mailing of the international search report 26 MAR 2003
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer Peggy Harod Telephone No. (703) 305-9703

Form PCT/ISA/210 (second sheet) (July 1998)*